Towards a grapho-phonologically parsed corpus of medieval Scots: Database design and technical solutions[1]

Joanna Kopaczyk, Benjamin Molineaux, Vasilios Karaiskos, Rhona Alcorn, Bettelou Los and Warren Maguire

1. Introduction

Automatic processing of digitized linguistic data, drawn from a wide range of available corpora, has become a staple methodology in the study of current and historical language use (for the aggregated corpora hub, see CoRD; see also Claridge 2008, Kytö 2012). Historical corpus linguistic methods have so far typically focussed on identification and labelling of units on higher levels of linguistic analysis, such as morphology and syntax (e.g. Santorini 2010), semantics (see the SAMUELS Project) and pragmatics (e.g. Archer and Culpeper 2003). These approaches have been considered better suited to early modern and later texts because by then the amount of variation, at least in written language, had already been subject to a large degree of standardisation, especially among major European languages which are best represented in corpus resources. Linguistic standardisation, in turn, makes automatic operations more straightforward.[2] That said, the application of these same corpus methodologies to historical phonology has met with important scepticism. For instance, in her review of historical corpus linguistic studies, Anne Curzan notes that a suitable corpus for phonological inquiry "would need to include editions that not only do not normalize the text but also preserve even the most idiosyncratic orthographic features, as these spellings can be evidence of pronunciation variants" (2008: 1097).

The project *From Inglis to Scots: Mapping sounds to spellings* (FITS) makes a first attempt at systematically exploiting the extensive – and often idiosyncratic – spelling variation attested in pre-modern non-standard spelling systems. The texts under scrutiny come from administrative and legal documents composed in various locations throughout Lowland Scotland between 1380 and 1500 in a variety that can be placed on a continuum of Germanic dialects spoken in the north of Britain. This variety, originally referred to as *Inglis*, came to be known as *Scottis* 'Scots' during the period of our investigation and flourished as a multi-purpose means of communication in medieval Scotland.[3] The material we employ in our analysis of spelling variants amounts to c.1,250 texts (c.0.4mln words) diplomatically transcribed from manuscripts and semantico-gramatically tagged for inclusion in the *Linguistic Atlas of Older Scots* (LAOS, Williamson comp. 2008).

The FITS project aims to achieve a systematic analysis of the relationships between spelling variants and the sound systems that underlie them, focusing on root-

[1] The authors gratefully acknowledge the financial support of the Arts and Humanities Research Council (AHRC grant number AH/L004542/1).

[2] Since standardisation of orthography was a gradual process, early and late modern texts still contain some spelling variation. To enable automatic processing of early modern texts, artificial standardisation can be carried out with, for instance, Variant Detector (VARD), developed at Lancaster University (Baron and Rayson 2008).

[3] For more information on the history of Scots, see McClure (1992) and chapters in Jones ed. (1997). For a diachronic outlook on Scots vowels, see Aitken and Macafee (2002). On the origins of Scots, as seen through the historical development of a selected phonological segment, see Alcorn *et al*. 2017.

morphemes of Germanic origin. The user of the freely available, fully searchable online database produced by the project will be able to find answers to questions such as:

- What sound(s) did the digraph <ch> represent in 15th-century Scots?
- When and where is theta-hardening ([θ] > [t]) attested in early Scots spellings?
- What are the reflexes of Old English /f/ in Scots?

The process of data entry has been carefully planned out with exactly such questions in mind. This paper presents the methodological and technical decisions undertaken in order to construct a database of synchronic relationships between orthographic variants and underlying sounds.[4] We begin by defining the grapho-phonological unit, and placing it within the theoretical literature on historical sound-spelling mappings. After this, we introduce the concept of grapho-phonological parsing, i.e. the resolution of individual word forms into sequences of sound values, which are then recorded in the FITS database. We then describe the design of the database and of its data-entry form, developed to record certain contextual information to assist with the interpretation of the synchronic relationships we capture. The resulting aggregated data, we will show, enables us to recover complex relationships between specific spelling choices and their most plausible sound values in any given phonotactic, morphological or lexical context. To illustrate the solutions adopted, we consider the sound-to-spelling mappings of two units in our corpus: <ch> and its associated sound values, and [ʃ] and its associated graphemes.

2. Methods and theoretical underpinnings

2.1 A grapho-phonological unit: Layers of interpretation

The framework within which we analyse spelling variation and its phonic significance is based around the taxonomic notion of the *littera* (Laing 1999, Laing and Lass 2003, 2005, 2009; Lass and Laing 2010, 2012).[5] This concept originates in an antique and medieval understanding of writing systems (Donatus, *Ars Maior*), which brings together several layers of interpretation of the "markings" on the page. We assume that these "markings" were made by scribes "capable of sophisticated and subtle linguistic analysis" (Laing and Lass 2003: 258), so we expect there to be a systematic connection – albeit not necessarily a one-to-one match – between orthographic choices and the underlying sound system(s). To situate the principles of the medieval notion of the *littera* in modern linguistics terms, we prefer to conceptualize it as a constellation of relationships involving: a grapheme, a set of one or more sounds, and a set of written manifestations (allographs). Two such constellations can be gleaned from Figure 1: one consisting of the relationship between the grapheme <ȝ> ('yogh') as it appears in *ȝer(e)* 'year' and *ȝerli* 'yearly' and the sound value [j], and the other between the same grapheme and the sound value [z] in *l(ett)reȝ* 'letters'.

[4] The FITS project also seeks to outline the diachronic dimension of sound change, as reflected in variant spellings. This aspect falls outside the scope of the present paper.

[5] For a detailed discussion of the *littera* as applied to medieval English manuscripts, see especially Laing (1999) and Laing and Lass (2003). These authors draw inspiration from Benediktsson's (1972) discussion of medieval orthographic theory and use the notation developed by Benskin (1997) for *litterae* (graphemes), *potestates* (sound values) and *figurae* (allographs).

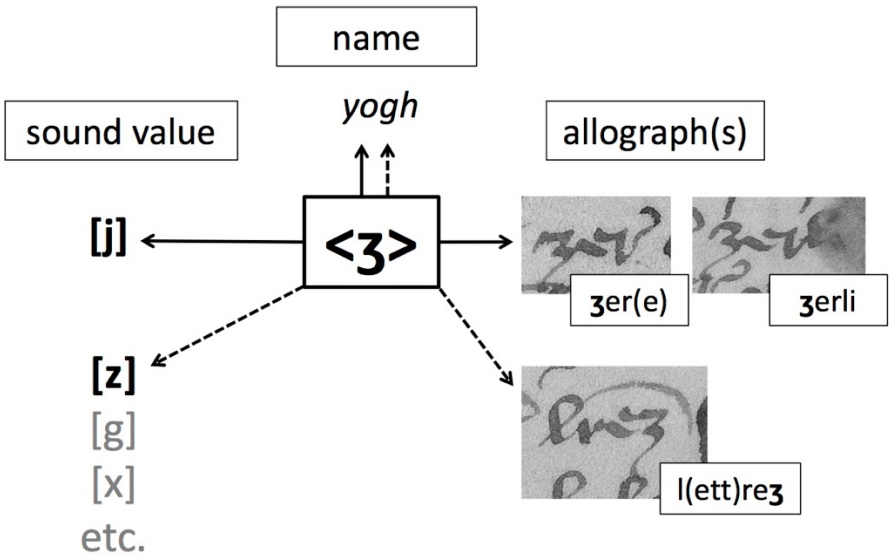

Figure 1. The components of a grapho-phonological unit in context (tokens from LAOS text #36, Laing Charters 805, 1447, Edinburgh)

2.2 Grapho-phonological parsing

The FITS project resolves each variant spelling through *grapho-phonological parsing*. Essentially, the procedure involves breaking up each form into a sequence of graphemic units and assigning each unit a suggested underlying sound value on the basis of:

(1) spelling variation across the corpus,
(2) what we know about the sounds of mediaeval Scots (e.g. based on Aitken and Macafee 2002 and Johnston 1997),
(3) what we know about the sounds of the preceding and following stages of the language,
(4) general theories of sound change and language change.

We thus make informed judgements in the same way as similar projects such as the *Linguistic Atlas of Early Medieval English*, that is to say: "our reconstructions are well enough supported so that if a responsible phonetician equipped with a time machine were able to hear the items represented, the symbol in question would be a reasonable transcriptional response" (Lass and Laing 2013: §2.4.2). Following Laing and Lass (2003: 268), we use the square brackets to represent "poorly resolved broad phonetic realizations", which is a consequence of historical phonology having orthography, rather than acoustic data, as its point of departure.

We thus establish a network of relationships between the graphemic units attested in our corpus and their plausible underlying sounds, starting with individual tokens to establish patterns across the entire data set. As Figure 1 shows, graphemes and sound values need not display a one-to-one match; indeed, they seldom do (see e.g. Venezky 1967, Bann and Corbett 2015, Alcorn *et al*. 2017). More often than not we are faced with substitution sets (Laing 1999, Laing and Lass 2003: 259-260, 262-263).[6] A

[6] Laing and Lass (2003) develop the notions of Litteral, Potestatic and Figural Substitution Sets (LSS, PSS and FSS). Our sets are based on the same principle but our nomenclature is rooted in modern linguistic concepts.

sound substitution set (SS set) identifies each of the sounds we associate with a specific grapheme (see §3.1), while a graphemic substitution set (GS set) identifies each of the graphemes we associate with a particular sound (see §3.2).[7] The challenge lies in the fact that many sounds and graphemes belong to more than one set. For instance, <ȝ > is a member of the GS set for [j], as in <ȝ ere> 'year', but it also belongs to the GS set for [x], as in <burȝ > 'burgh'. <i>, on the other hand, belongs to the GS set for [j], as in <iere> 'year', but not to the GS set for [x]. In the FITS capture tool, we provide ways to reconstruct GS and SS sets and establish the types and extent of such overlaps.

2.3 FITS: A database perspective

Using a static version of LAOS we organise our analysis in a relational database. A schematic diagram of the database structure is shown in Fig. 2, where each box corresponds to a database table.

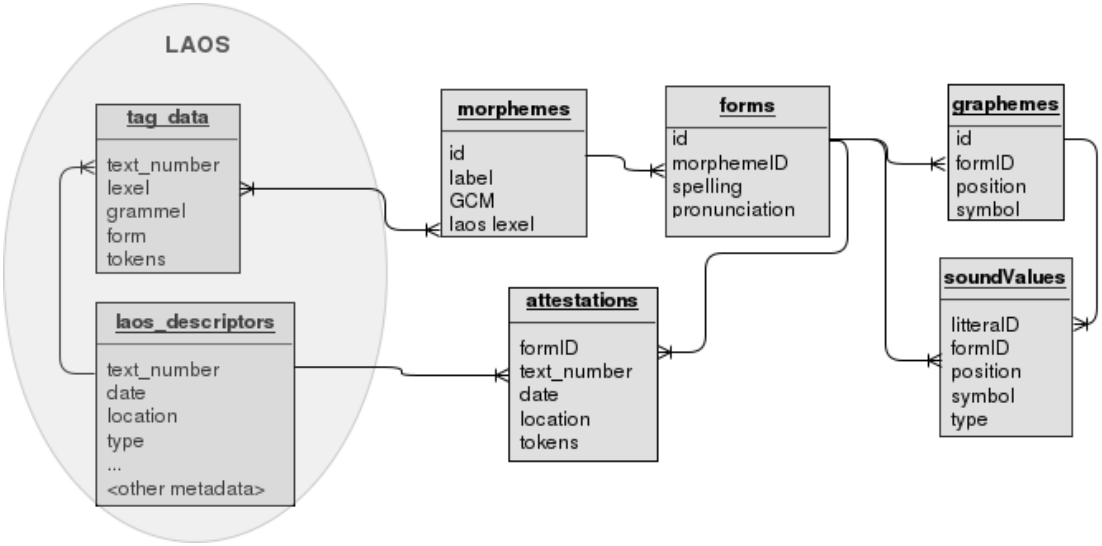

Figure 2. Database diagram of the grapho-phonological analysis for FITS

The LAOS data (see the shaded circle in Fig. 2) is equipped with various pieces of information required for the production of its associated linguistic atlas. The FITS data is derived from LAOS and is organised in a semi-hierarchical fashion, from *morphemes*, to *forms* (morpheme spellings and their attestations), to the *graphemes* and *sound values* that these spellings are resolved into. The basic unit of the FITS analysis is the root morpheme, indicated by braces, i.e. { }, and identified by its *label*,[8] its *grammatical context marker* (GCM),[9] and its corresponding *LAOS lexel*[10] (Table 1).

---

[7] For palaeographic investigations into the actual shapes of graphemes in a given scribal hand, we propose allographic substitution sets (AS sets) (see Fig. 1), although the study of allographs falls outside the FITS project.

[8] In general, the label for a given Scots morpheme is its present-day English cognate, partly because we expect most users to be familiar with Standard English, and partly because present-day reflexes of 15C Scots morphemes may be lacking or have no standard spelling. Where no present-day English equivalent is available, the Scots form is used as a label, using its citation form in the *Dictionary of the Scots Language*.

[9] As FITS needs less detailed grammatical information than a dialectal atlas, we have compacted 1,143 LAOS grammatical labels ('grammels') into 69 GCMs to describe the particular grammatical context of each token. In Table 1: *vps11*=1sg present, *vpt*=past tense, *vpsp*=present participle, *n*=noun.

[10] A lexel is a present-day gloss for a form encountered in LAOS texts.

Table 1. FITS morpheme {give} (selected records)

label	GCM	LAOS lexel
give	vps11	give
give	vpt	give
give	vpsp	give
give	vpsp	forgive
give	n	forgive

The attested spelling variants of a morpheme are the 'FITS *forms*'. Table 2 shows some of the attested forms of {give} in first person singular present tense contexts, the first context listed in Table 1. Each row of Table 2 records a spelling variant of the morpheme along with a link to the morpheme's unique identifier. The fields *spelling* and *pronunciation* correspond, respectively, to the LAOS transcription of the manuscript form and to the reconstruction determined by the grapho-phonological parsing procedure (§2.2). Although we provide sound values only for roots, we record the form of any pre- and post-root material (see §2.4) to allow for a systematic interpretation of their influence on the sound substance of the root.

Table 2. FITS forms of the morpheme {give}/vps11

spelling < >	pronunciation [ ]	morpheme ID
gyf	gɪf	1630
giffis	gɪf-#infl	1630
geve	gɐ:v-e	1630

We then pool together the number of spelling occurrences, date, location and text-type data from the LAOS material, and link that data to an attested form. This is done in the table *attestations* (Fig. 2).[11] The database records for the attestations of the form <gyf> from Table 2 are shown in Table 3.

Table 3. FITS attestations of the form <gyf>

form ID	LAOS transcription	text ID	text type	tokens	year	location
5720	GYF	364	charter / quitclaim	1	1425	395 830
5720	GYF	1236	book / record / burgh / court	1	1463	323 718
5720	GYF	1284	book / record / burgh / court	1	1468	323 718

[11] Data duplication in the attestations table was introduced to allow for separating the ultimate data source (LAOS) from the FITS-related records.

Each row links a FITS form (*form ID*) with the text in which it is attested (*text ID*), and shows the number of occurrences in that text (*tokens*). Some textual metadata (*text type, year, location*),[12] as well as the original LAOS transcription complete each record.

2.4 The FITS capture tool

In order to enter, aggregate, validate and visually inspect data we have designed a custom data capture tool. This is a collection of files and scripts created in HTML5, Javascript, and CSS for the front end, and in PHP for the back-end connection with the database. Fig. 3 shows how metadata from Tables 1-3 is represented in the interface.

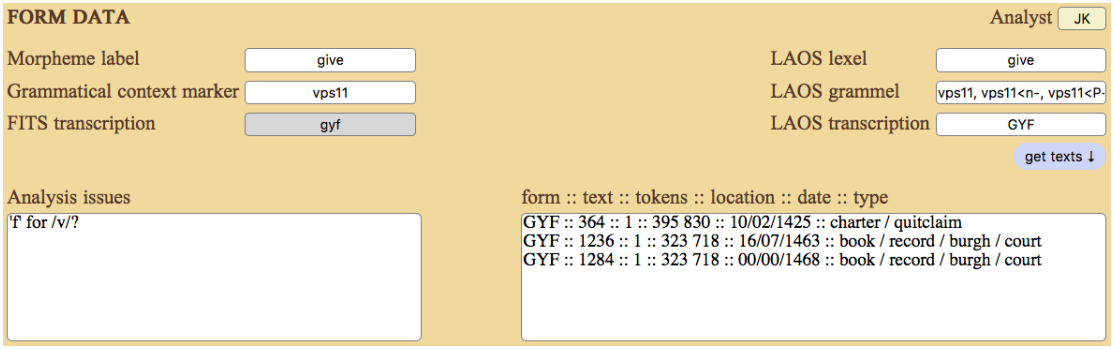

Figure 3: FITS capture tool screenshot for FITS form <gyf>

We use the same tool for the grapho-phonological parsing procedure. By recording segmentation separately for spellings and for sounds, we keep both levels of analysis self-contained without losing any interconnectivity. The compilers manually link each sound/grapheme pair and label it either as a stressed vowel (N=nucleus), an unstressed vowel (V), or a consonant (C). In a FITS form such as <yyffyn~>, a past participle form of {give}, the root is divided into three graphemic units (Fig. 4). The initial <y> is interpreted as a C, with an associated consonantal SS set (see §2.2) (including, but not limited, to [g], [ð], [θ], [j]),[13] while the second <y> has a N sound-repertoire (including, but not limited, to [ɪ], [ɛ], [ø]). Thus we are able to retrieve different sub-sets of sounds for a given grapheme, depending on their position in the root.

The third token of <y> in <yyffyn~> does not fall within the root morpheme, so we do not assign it a sound value (if it were a root element, it would have the V repertoire). However, it does get recorded as part of the morphological context in the "trailer" (T) slot. The "leader" (L) and "trailer" (T) slots on the edges of the morpheme (Figs. 4 and 5) capture its morphological and grapho-phonemic contexts, including the presence of trailing abbreviations and final <e>s.[14] In this case, the L-slot is empty, while the T-slot contains the inflectional *(#infl)* suffix <yn~>.[15] Note that the content of the T slot may influence the interpretation of the root. Here, the root-final graphemic unit <ff> is interesting as it implies a voiceless sound in an intervocalic context, i.e. where a voiced sound would be expected, historically. The presence of the inflectional

[12] Locations in the FITS database are represented by the first three digits of the northings and eastings for each text, as provided by LAOS. Examples of the use of FITS spatio-temporal data to investigate a linguistic feature can be found in Molineaux *et al*. 2016.

[13] As the corpus is under development and evolves dynamically as more forms are analysed, our SS and GS sets are not yet necessarily complete.

[14] There are three interpretations of final <e> in 15th-century Scots: (i) a residual schwa in final positions, which is very unlikely by this period unless intended as an archaism (Aitken and Macafee 2002: 69-71); (ii) a diacritic of some kind, most typically a length-marker for the root vowel; (iii) an otiose element without phonological consequence.

[15] In this case, the "grapho-phonemic" dropdown is empty.

material in the T-slot allows us to postulate that Older Scots word-final devoicing of voiced fricatives had, at least in some cases, spread (probably by analogy) into stem-final, pre-inflectional position (see Johnston, 1997:104, Maguire *et al*., in preparation).

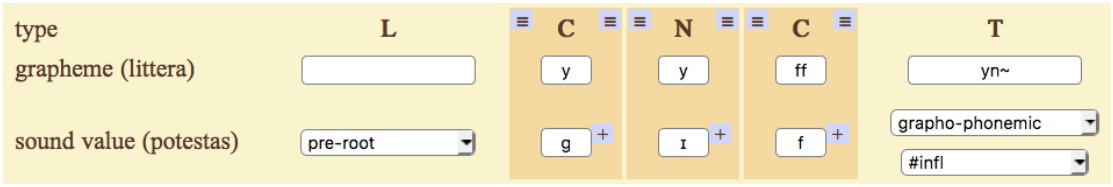

Figure 4: Postulated grapheme-to-sound-value equivalences for the form <yyffyn~> 'given' in the FITS capture tool

Figure 5 presents the analysis of a form of {worth} which uses all the fields in the FITS capture tool. The form occurs in <pe(n)nyworch~es>, a plural form of the compound PENNYWORTH. We start by segmenting the root form (<worch>) into its component graphemes, which we then classify by type, i.e. as N, V or C. The root is preceded by another root, {penny}, which we record and classify in L, and followed by a backwards-curving horizontal stroke and plural inflection, which we record and classify in T. The floating frame at the bottom is a custom keyboard for inserting special characters. Further comments on parsing are permitted in the "Analysis issues" free-text box. In this case, an unusual use of <ch> has prompted a comment which we unpack in our discussion of spelling-to-sound mappings below.

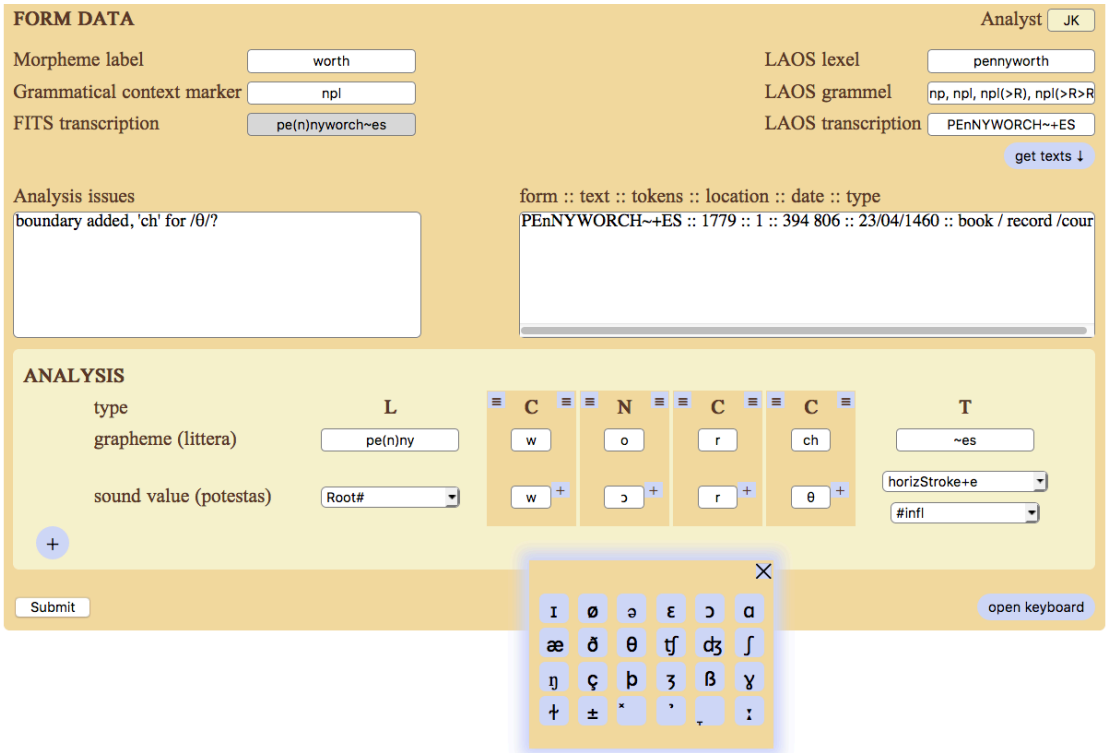

Figure 5: FITS capture tool interface with the full analysis of the form <pe(n)nyworch~es>, FITS morpheme {worth}

Once the analysis is complete and the data is submitted, the tool first validates the data to ensure, for example, that every grapheme has a corresponding sound value, and that no FITS form is saved in the database without its text attestation details (cf. Fig. 3). Following that, the capture tool collects all the data from the web form and either

creates new records in the database, or identifies previously entered records and updates those accordingly.

3. Multi-directional mapping

3.1 Spelling-to-sound mappings

Our corpus-based approach allows us to establish, quantify and visualise relations between units of sound and their spellings. For example, Figure 6 presents the SS set for the digraph <ch> in the FITS front-end grapho-phonemic visualisation tool. The floating box lists each sound associated with <ch> in our analyses and quantifies their frequencies. The visualisation to its left also identifies these associations, but within the entire network of correspondences in our database, and represents relative frequencies by the thickness of the connecting lines (light grey for spelling-to-sound and dark grey for sound-to-spelling).

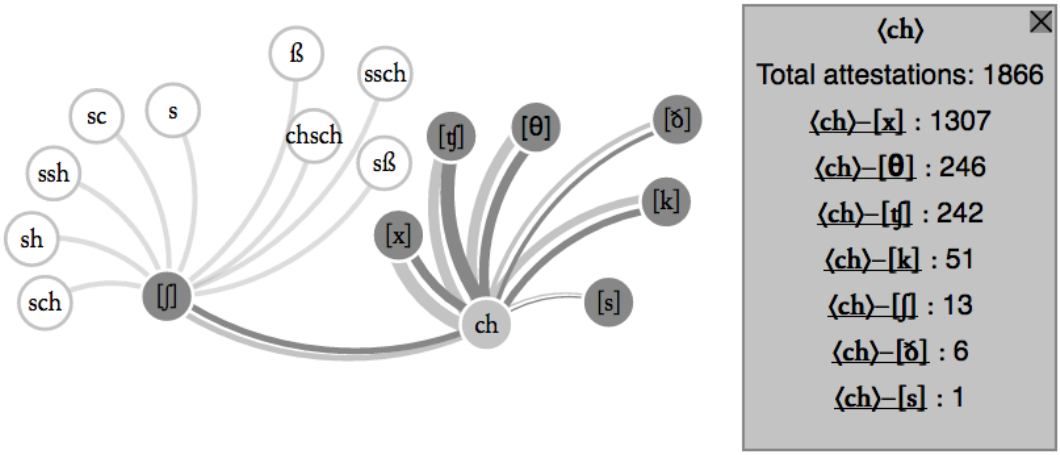

Figure 6: Visualisation of a sound substitution set (SS) for <ch> in the FITS database

Figure 6 tells us that the <ch> digraph is most frequently employed to represent [x], as in <richt> 'right' (< OE *riht*). The inclusion of [x] as part of the SS set for <ch> is supported by its continued use in present-day Scots dialects, and by the use of <ch> for [x] in non-Germanic words, e.g. in fifteenth-century forms of Gaelic *loch*. The second most frequent value of <ch> is [θ], as in <worch> (cf. Fig. 5). Use of <ch> for [θ] may be due to similarity, in some scribal systems, between the shapes of <t> and <c> on the page, such that the two became interchangeable in certain positions. Front-end users will also be able to investigate less frequent mappings throughout the database.

3.2 Sound-to-spelling mappings

The mirror image of a SS set, a GS set, can also be produced following these principles. As an example, consider a GS set for [ʃ] in Figure 7.

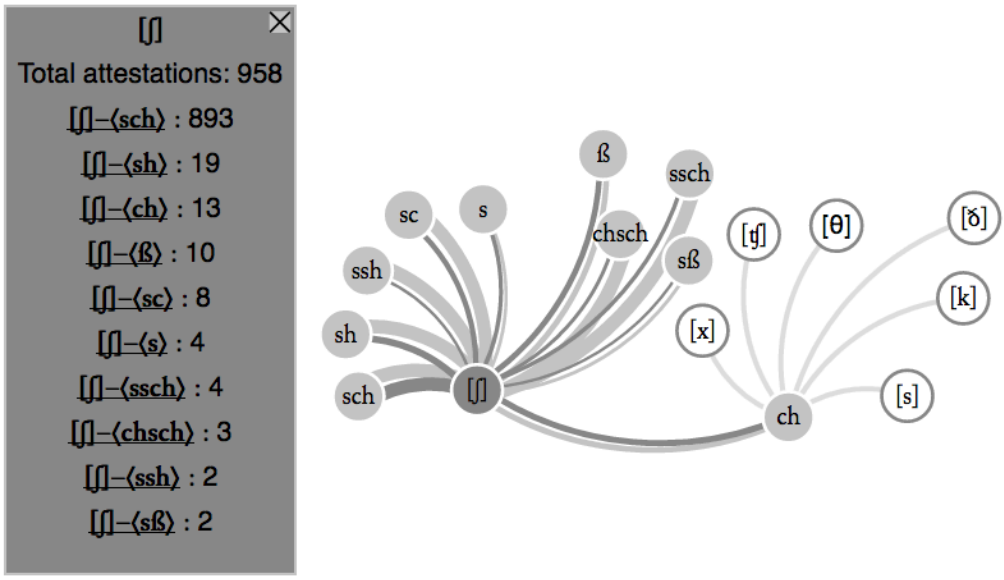

Figure 7: A graphemic substitution set (GS) for [ʃ] in the FITS database

The overwhelmingly preferred representation of [ʃ] as <sch> may be interpreted as a sign of incipient spelling standardisation in fifteenth-century Scots. The <sh> spelling, on the other hand, may be argued to carry hints of Southern English influence (Kniezsa 1997: 40), and competes with either <sch>, e.g. in {shilling}, or with <s>, e.g. in {sheriff} and {english}. We presume variation between <sch> and <sh> to have no phonological consequence, both graphemic sequences representing [ʃ], but variation between <sch> and <s> may also indicate variation in the sound substance, i.e. between [ʃ] and [s] (see Johnston 1997: 105). Here we follow the etymological category of the sound [ʃ], although further research into the matter will be required.

As the database grows, the grapho-phonological display tool changes dynamically, allowing the compilers to reassess the likelihood of individual mappings within a SS set (Figs. 6 and 7). The front-end user will be able to examine and re-evaluate proposed GS and SS sets for the whole corpus and for individual texts, locations or time spans.

4. Conclusions

The resources produced by the FITS project will enable users to trace orthographic and phonological developments – and their interactions – within the scope of the corpus, thus furthering our understanding of the features of the earliest attested Scots language. Key to this process is allowing end-users to interact dynamically with the data by:

- selecting specific sound, orthographic and grammatical environments,
- defining temporal and spatial domains for search results,
- tracing etymological sources morpheme-by-morpheme and sound-by-sound via a Corpus of Changes (under development),
- linking morphemes to associated entries in the online *Dictionary of the Scots Language* and *Oxford English Dictionary*,
- and accessing full texts.

FITS analyses are open to interpretation and re-evaluation, so we aim to provide users with a downloadable version of the complete database along with the relevant documentation. Other outputs and visualisations will depend on the kinds of questions that users would like to ask independently of FITS, and on how they need to access the data, for example as maps, graphs or tables.

The introduction of the grapho-phonological parsing procedure is a major advance in the development of corpus linguistics, especially in the context of historical records. It facilitates the systematic investigation of spelling systems and their underlying phonological substance. At present, the process of data entry is still manual and labour-intensive but what we learn from it will help to develop automated solutions on a larger scale. The way we approach our data should be applicable to any historical sound system that pre-dates the standardisation of its spelling; that is, following the uniformitarian principle, we can expect that scribes and printers of other vernaculars also used structured systems with multilayered mappings, so the tool design and the general principles of FITS should be applicable in these other contexts.[16] In fact, the principles of grapho-phonological parsing ought to allow for a fresh look at uncodified orthographic systems representing indigenous and minority languages as well regional and social dialects today. A good testing ground would be the growing body of non-standard spellings in online communication which seem to be produced ad-hoc but are far from random. Ultimately, the methods we have presented here are a novel way of finding the underlying systematicity and variation across sound systems as filtered through the medium of spelling practices.

References

Aitken, A. J., & Macafee, C. 2002. *The Older Scots vowels: A history of the stressed vowels of Older Scots from the beginnings to the eighteenth century*. Edinburgh: Scottish Text Society.

Alcorn, R., B. Molineaux, J. Kopaczyk, V. Karaiskos, B. Los and W. Maguire. 2017. 'The emergence of Scots: Clues from Germanic *a reflexes' in J. Cruickshank and R. McColl Millar (eds.) *Before the Storm: Papers from the Forum for Research on the Languages of Scotland and Ulster triennial meeting, Ayr 2015*, pp. 1-32. Aberdeen: Forum for Research on the Languages of Scotland and Ulster.

Archer, D. and J. Culpeper. 2003. 'Sociopragmatic annotation: New directions and possibilities in historical corpus linguistics' in A. Wilson, P. Rayson and T. McEnery (eds.) *Corpus linguistics by the lune: A festschrift for Geoffrey Leech*, pp. 37-58. Frankfurt a. Main: Peter Lang.

Bann, J. and J. Corbett. 2015. *Spelling Scots. The orthography of literary Scots 1700-2000*. Edinburgh: Edinburgh University Press.

Baron, A. and P. Rayson. 2008. 'VARD 2: A tool for dealing with spelling variation in historical corpora' in Proceedings of the Postgraduate Conference in Corpus Linguistics, Aston University, Birmingham, UK, 22 May 2008.

Benediktsson, H. 1963. 'Some aspects of Nordic umlaut and breaking', *Language*, 39(3), pp. 409-431.

Benskin, M. 1997. 'Texts from an English township in late medieval Ireland', *Collegium Medievale* 1-2, pp.91–170.

Claridge, C. 2008. 'Historical corpora' in A. Lüdeling and M. Kytö (eds.) *Corpus linguistics: An international handbook*. Vol. 1, pp. 242-259. Berlin: Mouton de Gruyter.

CoRD = Corpus Resource Database, http://www.helsinki.fi/varieng/CoRD/

[16] One of our co-authors has received British Academy funding to adapt the FITS software to undertake a pilot study of early Middle English spelling systems using data from *A Linguistic Atlas of Early Middle English*, see further: www.amc.lel.ed.ac.uk/?page_id=1692.

Curzan, A. 2008. 'Historical corpus linguistics and evidence of language change', in A. Lüdeling and M. Kytö (eds.) *Corpus linguistics: An international handbook*. Vol. 1, pp. 1091-1109. Berlin: Mouton de Gruyter.

Dictionary of the Scots Language, Scottish Language Dictionaries, http://www.dsl.ac.uk/ http://www.gla.ac.uk/schools/critical/research/fundedresearchprojects/samuels/

Johnston, P. 1997. 'Older Scots phonology and its regional variation', in C. Jones (ed.), *The Edinburgh history of the Scots language*, pp. 47-111. Edinburgh: Edinburgh University Press.

Jones, C. (ed.), 1997. *The Edinburgh history of the Scots language*. Edinburgh: Edinburgh University Press.

Kniezsa, V. 1997. 'The origins of Older Scots orthography', in Ch. Jones (ed.) *The Edinburgh history of the Scots language*, pp. 24-46. Edinburgh: Edinburgh University Press.

Kytö, M. 2012. 'New perspectives, theories and methods: Corpus linguistics', in A. Bergs and L. J. Brinton (eds.) *English historical linguistics. An international handbook*. Vol. 2, pp. 1509-1531. Berlin: Mouton de Gruyter.

Laing, M. and R. Lass. 2003. 'Tales of 1001 nists: The phonological implications of litteral substitution sets in some thirteenth-century South-West Midland texts', *English Language and Linguistics* 7(2), pp.257-278.

Laing, M. and R. Lass. 2005. 'Early Middle English knight: (Pseudo)metathesis and lexical specificity', *Neuphilologische Mitteilungen* 106(4), pp. 405-423.

Laing, M. and R. Lass. 2009. 'Shape-shifting, sound-change and the genesis of prodigal writing systems', *English Language and Linguistics* 13(1), 1-31.

Laing, M. 1999. 'Confusion wrs confounded: Litteral Substitution Sets in Early Middle English writing systems', *Neuphilologische Mitteilungen* 100(3), 251-270.

Lass, R. and M. Laing. 2010. 'In celebration of Early Middle English 'h'', *Neuphilologische Mitteilungen* 111(3), 345-354.

Lass, R. and M. Laing. 2012. ''ea' in early Middle English: from diphthong to digraph' in D. Denison, R. Bermúdez-Otero, C. McCully and E. Moore (eds.) *Analysing older English*, pp. 75-118. Cambridge: Cambridge University Press.

Lass, R. and M. Laing. 2013. 'Introduction, Chapter 2' in *A Linguistic Atlas of Early Middle English, 1150–1325*, Version 3.2, comp. by M. Laing [http://www.lel.ed.ac.uk/ihd/laeme2/laeme2.html]. Edinburgh: © The University of Edinburgh.

Maguire, W. *et al*. In prep. Investigating evidence for final [v]-devoicing in Older Scots.

Molineaux, B., J. Kopaczyk, W. Maguire, R. Alcorn, V. Karaiskos and B. Los. 2016. 'Tracing L-vocalisation in early Scots'. *Papers in Historical Phonology* 1, pp. 187-217.

McClure, J. D. 1994. 'English in Scotland' in R. W. Burchfield (ed.), *The Cambridge History of the English Language*. Vol. 5: *English in Britain and Overseas*, pp. 23-93. Cambridge: Cambridge University Press.

SAMUELS Project = Semantic Annotation and Mark-Up for Enhancing Lexical Searches, University of Glasgow. http://www.gla.ac.uk/schools/critical/research/fundedresearchprojects/samuels/#/academicprojectteam,projectoutputs

Santorini, B. 2010. Annotation manual for the PPCME2, PPCEME and PCEEC. Release 2. http://www.ling.upenn.edu/hist-corpora/

Venezky, R. L. 1967. 'English orthography: Its graphical structure and its relation to sound', *Reading Research Quarterly* 2(3), pp. 75-105.

Williamson, K. 2008. *LAOS: A Linguistic Atlas of Older Scots, Phase 1: 1380-1500*. Retrieved from http://www.lel.ed.ac.uk/ihd/laos1/laos1.html. The University of Edinburgh.